Normal distribution Practice problem

1. Generate 200 normally distributed data using Python with any values for location and scale parameters. Verify that the empirical rule for your data.

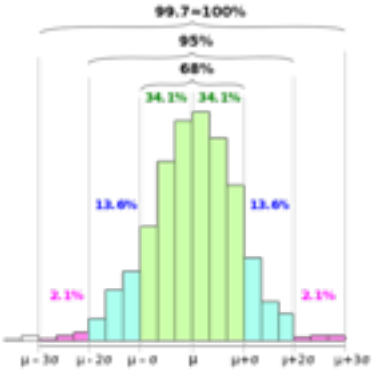
Empirical Rule for symmetric distributions: The empirical rule says that, in a normal/approximately normal data set, virtually every piece of data will fall within three standard deviations of the mean. The mean is the average of all the numbers within the set.

The empirical rule is also referred to as the Three Sigma Rule or the 68-95-99.7 Rule because:

Within the first standard deviation from the mean, 68% of all data rests

95% of all the data will fall within two standard deviations

Nearly all the data – 99.7% – falls within three standard deviations (the .3% that remains is used to account for outliers, which exist in almost every dataset)



2. The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

You can find the corresponding data on canvas under the modules named “Probability Distributions”. There is a file named description.txt to explain the variables in the data. Primary data is named as “train.csv”.

Complete the following tasks and explain your observations:

1. Check the ‘SalePrice’ variable. Find the descriptive statistics of the variable.
2. Plot the histogram of the ‘SalePrice’ and the density plot. Do you think ‘SalePrice’ follows normal distribution? How can you verify that?
3. Are there any missing values for ‘SalePrice’? If any, please remove those missing values.
4. Are there any outliers for ‘SalePrice’? Remove those as well. Do you think removal of outliers was helpful in assuming normal distribution for the variable?

5. Do you think standardizing this variable will more useful than working with the original data? Justify your answer.
6. Try to draw the histogram and density plot again. Is there any change you see from the previous plot?
7. Take a logarithmic transformation of the data and check the normality assumptions again. Do you think the transformation made a relative difference from the previous observations?
8. Try parts 1 – 6 for the variable “GrLivArea” and comment on that. Does that process gives you similar results? Do you see any difference, if any?
9. Try parts 1 – 6 for the variable “TotalBsmtSF” and comment on that. Does that process gives you similar results? Do you see any difference, if any?